



中国研究生创新实践系列大赛
“华为杯”第十七届中国研究生数学建模竞赛

学　　校	中国科学技术大学
参赛队号	**20103580004**
队员姓名	**1.** 甄宇
	2. 张一波
	3. 程泽坤

中国研究生创新实践系列大赛
“华为杯”第十七届中国研究生
数学建模竞赛

题　　目　汽油辛烷值优化建模

摘　　要：

本文通过机器学习算法和数据挖掘技术，建立了汽油辛烷值（RON）损失的预测模型，在保证汽油产品脱硫效果的前提下，以降低汽油辛烷值损失在 30% 以上为目标，给出每个样本的优化操作条件。

针对第一问，对 285 号和 313 号原始数据用给定的数据处理规范进行预处理，对每一列变量求期望，得到最终的数据补充到相应的样本编号。

针对第二问，首先对数据样本按照最大最小限幅以及拉依达法则进行列变量清洗，通过两轮特征筛选确定变量：结合处理后的数据，针对产品中硫含量、RON 损失值用 lightGBM 做特征权重打分，筛选出权重排名较前、原料性质、待生吸附剂、再生吸附剂等 18 个变量。针对排名较低且独立出现的 57 个变量进行相关性分析，基于决策树模型二次选出 5 个变量。最终共选出 23 个主要变量，其中操作变量 17 个。

针对第三问，通过模型融合构建集成学习模型，分别对硫含量、RON 损失值预测。构建 4 个相关性较低的基学习器：多层感知机、随机森林、决策树、lightGBM。将 4 个基学习器进行模型融合，集成到元学习器 GBDT 中，实现 2 层集成学习模型。根据误差评价指标，集成模型的预测误差较小，有较好的预测效果。

针对第四问，以第三问构建的两个模型作为目标函数，采用基于 NSGA-II 的遗传算法进行多目标优化，找到了 325 个所有样本的帕累托最优解，并筛选出了 304 个样本，其优化后的操作变量取值可使得样本满足产品性质要求。对应最优的操作变量取值附成表格，以供后续实际参考。在硫含量小于 $5\mu g/g$ 的条件下，模型平均可以优化 RON 损失减小 47% 左右。

针对第五问，利用问题四中优化模型，寻找 133 号样本的帕累托最优解对应的最优操作变量。通过对比操作变量原始值和最优值，判定 $S-ZORB.PDT\_3002.DACA$、$S-ZORB.ZT\_2634.DACA$、$S-ZORB.FC\_5103.DACA$ 三个操作变量较为重要。对 133 号样本逐步调整形成的 15 个调整样本预测，结果显示随着操作变量趋于最优，产品中硫含量、RON 损失值逐渐降低，而 RON 损失值最大降幅达到 54.84%。

关键字： 特征筛选　集成学习　遗传算法　帕累托最优

目录

1. 问题重述

1.1 问题背景

汽油是小型车辆的主要燃料，汽油燃烧产生的尾气排放对大气环境有重要影响。为此，世界各国都制定了日益严格的汽油质量标准。汽油清洁化重点是降低汽油中的硫、烯烃含量，同时尽量保持其辛烷值。

辛烷值（以 RON 表示）是反映汽油燃烧性能的最重要指标，并作为汽油的商品牌号。现有技术在对催化裂化汽油进行脱硫和降烯烃过程中，普遍降低了汽油辛烷值。

化工过程的建模一般是通过数据关联或机理建模的方法来实现的，取得了一定的成果，但是由于炼油工艺的复杂性以及操作变量的高度非线性及相互强耦联，传统数据关联模型难以对过程优化作出及时响应，效果不佳。

现有某石化企业运行 4 年的催化裂化汽油精制脱硫装置并积累了大量历史数据，其汽油产品辛烷值损失平均为 1.37 个单位相较于同类装置的最小损失值 0.6 个单位有较大的优化空间。现通过数据挖掘技术来解决该场景化工过程建模问题。

1.2 问题的提出

根据从催化裂化汽油精制装置采集的 325 个数据样本（每个数据样本都有 354 个操作变量），通过数据挖掘技术来建立汽油辛烷值（RON）损失的预测模型，并给出每个样本的优化操作条件，在保证汽油产品脱硫效果（欧六和国六标准均为不大于 $10\mu g/g$，但为了给企业装置操作留有空间，本次建模要求产品硫含量不大于 $5\mu g/g$）的前提下，最终完成降低汽油辛烷含量损失降幅在 30% 以上。

现根据以上背景以及所提供数据完成以下任务：

（1）参考近四年的工业数据的数据样本，对 285 号和 313 号样本原始数据根据给定的样本处理方法对数据样本进行预处理，填入样本数据集中对应的数据样本编号中，以便进一步分析。

（2）数据样本提供了 325 个样本数据，以及建立辛烷损失值模型所需要的 367 个操作变量。通过降维的方法筛选出建模的主要变量，并给出详尽的分析。

（3）采用（1）（2）中完成的数据样本以及建模变量，使用数据挖掘技术建立辛烷值（RON）损失预测模型，并验证。

（4）在保证产品硫含量不大于 $5\mu g/g$ 的前提下，利用（3）中的模型对应的 325 个数据样本可操作变量进行优化，并给出辛烷值（RON）损失降幅大于 30% 的主要变量优化后的操作条件。

（5）对 133 号样本，图形展示（2）中选定的主要操作变量在优化调整过程中对应辛烷值和硫含量的变化轨迹。

2. 问题分析

2.1 针对问题一

本题要求对 285 号和 313 号原始数据根据指定的数据处理规范进行预处理并填入相应的样本编号。样本的原始数据每编号各 40 组，我们将每组的数据用给定的数据处理规范进行验证，得到最终的数据，最后对每一列求期望补充相应的样本编号。

2.2 针对问题二

本题首先对数据样本按照最大最小限幅以及拉依达法则进行列变量清洗。

第一轮筛选，将处理后的数据针对产品中硫含量、RON 损失值用 lightGBM 做特征权重打分并以一定规则筛选出权重排名较前以及原料性质附带待生吸附剂、再生吸附剂等 17 个变量。

第二轮筛选，针对排名较低且独立出现的 57 个变量进行相关性分析，针对相关性 >0.8 的两两变量避免同时出现。之后基于决策树模型再做特征，选择选出最终 5 个变量并整合第一轮 17 个变量，得到建模所需要的 22 个主要变量，具体流程如图1所示。

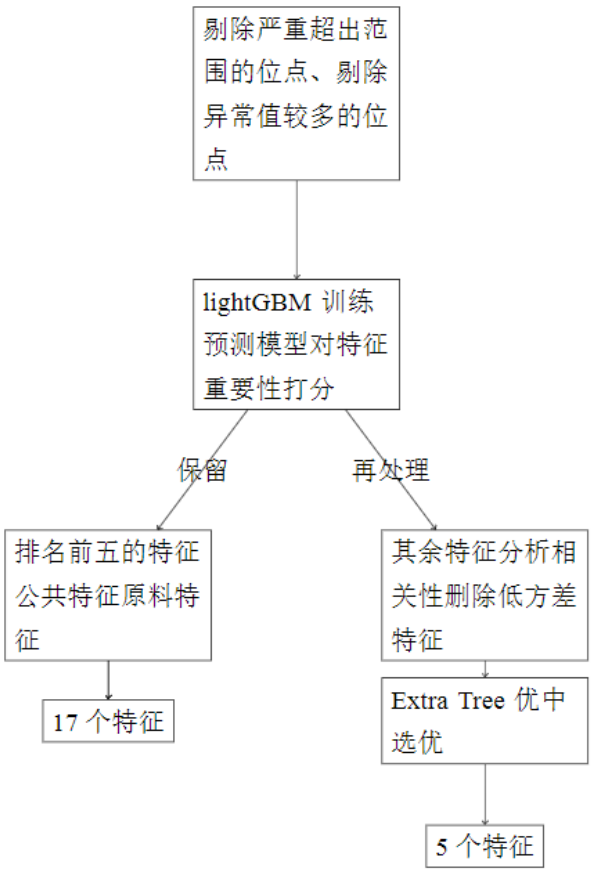


图 1　问题二分析求解流程图

2.3 针对问题三

本题利用模型融合的方法，构建集成学习模型，分别建立产品硫含量、RON 损失值预测的集成学习模型。

首先分别训练 4 个基学习器：多层感知机、随机森林、决策树、梯度提升决策树，构建 4 个相关性较低的弱学习器。

再将 4 个基学习器进行模型融合，集成到梯度下降树（GBDT）中，实现 2 层的集成学习模型，如图2所示。

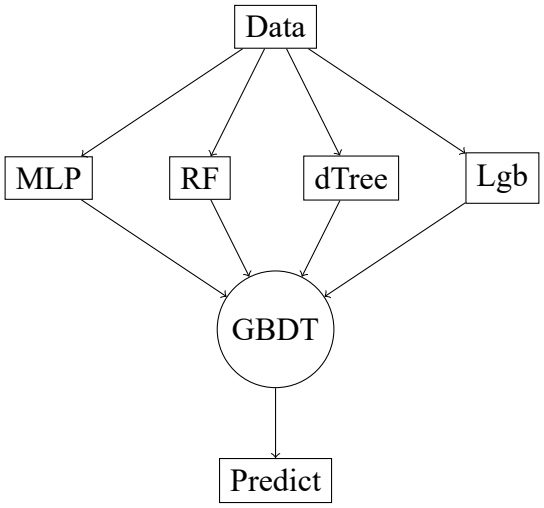


图 2　基学习器与元学习器融合

根据评价指标均方误差（MSE）、平均绝对误差（MAE）, 集成的模型对硫含量、RON 损失值的预测准确度远高于单一基学习器，且误差都较小。模型有较好的预测效果。

2.4 针对问题四

问题要求给出操作变量的优化方案，使得依据优化操作变量生成的 RON_{loss} 和 CP_S 满足给定要求。将此问题建模为多目标优化问题，目标函数即为第三问建立的两个模型。由遗传算法计算得到 325 个样本的操作变量的帕累托最优取值，再在其中进行筛选，删去不满足题设要求的帕累托最优解，即可找到完全满足题设要求的优化操作变量取值。

2.5 针对问题五

根据问题四中构建的优化策略，在保证优化目标的前提下，寻找 133 号样本点的帕累托最优解对应的最优操作变量。结合操作变量范围，将初始操作变量逐步调整至最优数值。每一步调整都会产生 1 33 号新的样本，利用预测模型对其预测，得到操作变量优化调整过程中对应的汽油辛烷值和硫含量的变化。

3. 符号说明

符号	意义
RON_{loss}	辛烷损失值
CP_S	产品含硫量
DS_S	待生吸附剂性质中的硫 $S.wt\%$
DS_C	待生吸附剂性质中的焦炭 $.wt\%$
ZS_S	再生吸附剂性质中的硫 $S.wt\%$
ZS_C	再生吸附剂性质中的焦炭 $.wt\%$
YL_{fang}	原料性质中的芳香烃
YL_{ting}	原料性质中的饱和烃
YL_{Br}	原料性质中的溴
YL_{RON}	原料性质中的辛烷值
$Model_{RON_{loss}}$	第三问分析后建立的辛烷损失值模型
$Model_{CP_S}$	第三问分析后建立的产品硫含量模型

4. 问题一的分析与求解

4.1 原始数据说明

285 号和 313 号样本原始数据，采集时间分别为 2017 年 7 月 17 日 6 时至 2017 年 7 月 17 日 8 时和 2017 年 5 月 15 日 6 时至 2017 年 5 月 15 日 8 时，数据采集频次为 3 分钟/次，每个样本各有 40 组数据，每组数据包括 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、2 个产品性质等变量以及另外 354 个操作变量，共计 367 个变量。

4.2 原始数据预处理

样本确定要求指出：以辛烷值数据测定的时间点为基准时间，取其前 2 个小时的操作变量数据的平均值作为对应辛烷值的操作变量数据。对 285 号和 313 号每一列数据分别求平均值，并将其填入样本数据集相应编号中。

5. 问题二的分析与求解

5.1 基础数据清洗

第一问将原始样本数据进行预处理并整合到了数据样本中。对数据样本的遍历检查过程中，观察到有部分操作变量不符合要求规范。

5.1.1 最大最小限幅

样本确定方法要求（4）指出：根据工艺要求与操作经验，总结出原始数据变量的操作范围。根据题目所给出的各个操作变量的具体范围，数据样本有部分数据编号对应的变量不在最大最小限幅内。

对数据样本每一列的最大最小值进行统计之后，得到列号为 22、49、83、106、111、238、323、343 的操作变量不符合最大最小限幅要求。其数据样本中的最大最小值与操作范围的最大最小值对比如表1所示。

表 1　实际最大最小值不在理论范围内的操作变量

列号	变量名称	变量单位	理论最小 | 最大值	实际最小 | 最大值
22	$S-ZORB.AT\_5201.PV$	$\mu g/g$	0|5	-0.3202|32.2342
49	$S-ZORB.SIS\_LT\_1001.PV$	%	40|80	190.545|1298500
83	$S-ZORB.TC\_2607.PV$	$°C$	450|520	447.486|516.211
106	$S-ZORB.FT\_1202.TOTAL$	无	-520000|175	-541103|150.818
111	$S-ZORB.FT\_1204.TOTAL$	无	45000|2500000	-49395.5|8397400
238	$S-ZORB.TE\_2001.DACA$	$°C$	-243600|12500000	-243602|12265800
323	$S-ZORB.AT-0012.DACA.PV$	无	0.2|2.2	0.1744|2.0371
343	$S-ZORB.CAL.LEVEL.PV$	无	4200|20000	-4269.46|19162

5.1.2 3σ 准则

拉依达准则（3σ 准则）常用于对异常数据进行处理。设对被测量变量进行等精度测量，得到 $x_1, x_2, ..., x_n$，算出其算术平均值 x 及剩余误差 $v_i = x_i - x(i = 1, 2, ..., n)$，并按贝塞尔公式算出标准误差 σ，若某个测量值 x_b 的剩余误差 $v_b(1 \leq b \leq n)$，满足 $|v_b| = |x_b - x| > 3\sigma$，则认为 x_b 是含有粗大误差值的坏值，应予剔除。

贝塞尔公式如下：

$$\sigma = (\frac{1}{n-1}\sum_{i=1}^{n} v_i^2)^{\frac{1}{2}} = (\frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1})^{\frac{1}{2}}$$

样本确定方法中（5）要求根据该法则对数据样本中整体的异常值进行处理。

根据法则公式对所有数据样本进行统计，得到统计如下，由于数据样本的操作变量较多，有较多数据样本的列满足拉依达法则的坏值预估，将存在异常值的数据样本输出并统计每列的异常数据数及其占总数据数的比例如表2所示。

表 2　异常数据较多的列以及占比情况

列号	操作变量名	异常数据数	异常数据占比%
329	$S-ZORB.TE\_6001.DACA.PV$	8	2.5
199	$S-ZORB.PDT\_3602.DACA$	7	2.2
168	$S-ZORB.TE\_5002.DACA$	7	2.2
302	$S-ZORB.TXE\_2203A.DACA$	6	1.8
170	$S-ZORB.FC\_5203.DACA$	6	1.8
202	$S-ZORB.SIS\_PT\_6007.PV$	6	1.8
244	$S-ZORB.PC\_2401.PIDA.SP$	6	1.8
175	$S-ZORB.FT\_2431.DACA$	6	1.8
177	$S-ZORB.TC\_2201.OP$	6	1.8
305	$S-ZORB.TE\_5009.DACA$	6	1.8

同时根据每一列异常值出现的数量绘制了异常值数量-出现频率直方图，如图3所示。

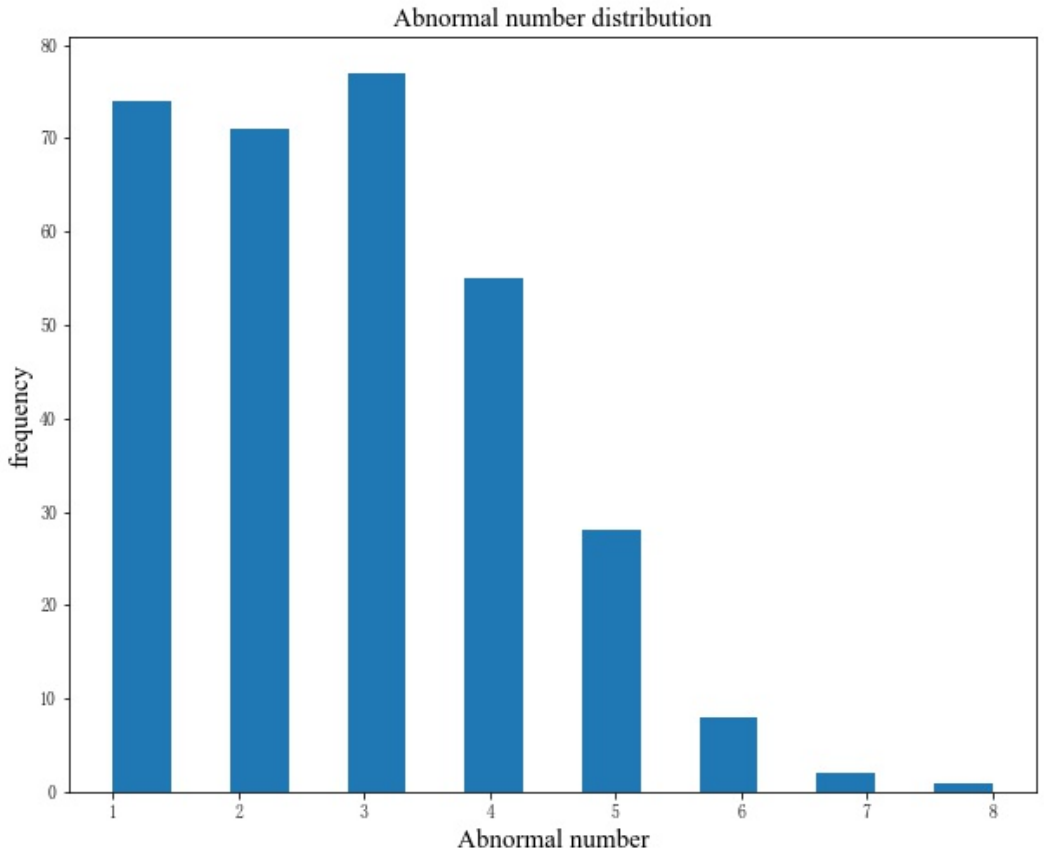


图 3　异常值数量 – 出现频率直方图

根据每一列出现的异常值，我们淘汰出现异常值较多（> 6）的操作变量。列号为 329、199、168 的操作变量被剔除。

根据最大最小限幅以及 3σ 准则，我们剔除了操作变量 $S-ZORB.AT\_5201.PV$（列号 22）、$S-ZORB.SIS\_LT\_1001.PV$（列号 49）、$S-ZORB.TC\_2607.PV$（列号 83）、$S-ZORB.FT\_1202.TOTAL$（列号 106）、$S-ZORB.FT\_1204.TOTAL$（列号 111）、$S-ZORB.TE\_5002.DACA$（列号 168）、$S-ZORB.PDT\_3602.DACA$（列号 199）、$S-ZORB.TE\_2001.DACA$（列号 238）、$S-ZORB.AT-0012.DACA.PV$（列号 323）、$S-ZORB.TE\_6001.DACA.PV$（列号 329）、以及 $S-ZORB.CAL.LEVEL.PV$（列号 343）对结果产生的影响。

5.2 特征筛选算法

在经过数据清洗之后，我们排除了异常情况较多的列。为了进一步筛选有代表性的操作变量并进行建模，进行特征筛选分析。

对于一个特定的数据集和学习算法来说，哪一个或哪些个特征的相关有效性是未知的。因此，需要根据数据样本从所有特征中选择出对于学习算法有益或更有代表性的相关特征。在实际应用中，操作变量过多会造成冗余（多组特征高度相关，造成性能时间的浪费），噪声（相关性不高的特征被用于建模，结果准确性下降）等问题。如果只选择所有特征中的部分特征构建模型，那么可以大大减少学习算法的运行时间，也可以增加模型的可解释性。

常用的特征筛选算法分类如下 [1, 2]：

5.2.1 Filter 方法（过滤式）

具体实现大体思想是先进行特征选择，然后去训练学习器，所以特征选择的过程与学习器无关。相当于先对特征进行过滤操作，然后用特征子集来训练分类器。

主要思想：对每一维特征“打分”，即给每一维的特征赋予权重，这样的权重就代表着该特征的重要性，然后依据权重排序。

主要方法：

- Chi-squared test（卡方检验）
- Information gain（信息增益）
- Correlation coefficient scores（相关系数）

其优势在于运行速度较快，应用广泛，但是无法提供反馈，特征选择的标准/规范的制定是在特征搜索算法中完成，学习算法无法向特征搜索算法传递对特征的需求。另外，可能处理某个特征时由于任意原因表示该特征不重要，但是该特征与其他特征结合起来则可能变得很重要。

5.2.2 Wrapper 方法（封装式）

对于样本数据集，直接把最后要使用的分类器作为特征选择的评价函数，对于特定的分类器选择最优的特征子集。主要思想：将子集的选择看作是一个搜索寻优问题，生成不同的组合，对组合进行评价，再与其他的组合进行比较。这样就将子集的选择看作是一个优化问题，这里有很多的优化算法可以解决，尤其是一些启发式的优化算法，如优化算法-粒子群算法、优化算法-人工蜂群算法等。

该种方法分类递归地消除特征，其优势在于对特征选择的标准/规范是在学习算法的需求中展开的，能够考虑学习算法所属的任意学习偏差，从而确定最佳子特征，真正关注的是学习问题本身。由于每次尝试针对特定子集时必须运行学习算法，所以能够关注到学习算法的学习偏差/归纳偏差，因此封装能够发挥巨大的作用。但其运行速度远低于过滤算法，实际应用没有过滤方法广。

5.2.3 Embedded 方法（嵌入式）

该种方法分类将特征选择嵌入到模型训练当中，其训练可能是相同的模型，但是特征选择完成后，还能给予特征选择完成的特征和模型训练出的超参数，再次训练优化。

主要思想：在模型既定的情况下学习出对提高模型准确性最好的特征。也就是在确定模型的过程中，挑选出那些对模型的训练有重要意义的特征。

主要方法：

- 用带有 L1 正则化的项完成特征选择（结合 L2 惩罚项进行优化）
- 随机森林平均不纯度减少法/平均精确度减少法。

该种方法优势在于能够考虑学习算法所属的任意学习偏差。训练模型的次数小于 Wrapper 方法，相对节省时间。但是运行速度仍然不占优势。

5.3 特征选择的实现

基于以上几种特征选择算法的介绍与分类，考虑本题背景和数据样本特性，我们的分析如下：

5.3.1 计算特征重要性

在使用 GBDT（全梯度下降树）、RF（随机森林）、XGBoost（极度梯度提升）等树类模型 [3] 建模时，可以在训练过程中通过记录特征的分裂总次数、总或平均信息增益来量化特征重要性。在训练过程中，可以使用特征在整个 GBDT。XGBoost 中使用的次数或带来的总/ 平均信息增益来给特征重要性打分，最后进行排序。

作为单个决策树模型，在模型建立时实际上是寻找到某个特征合适的分割点。这个信息可以作为衡量所有特征重要性的一个指标。

基本思路如图4所示：

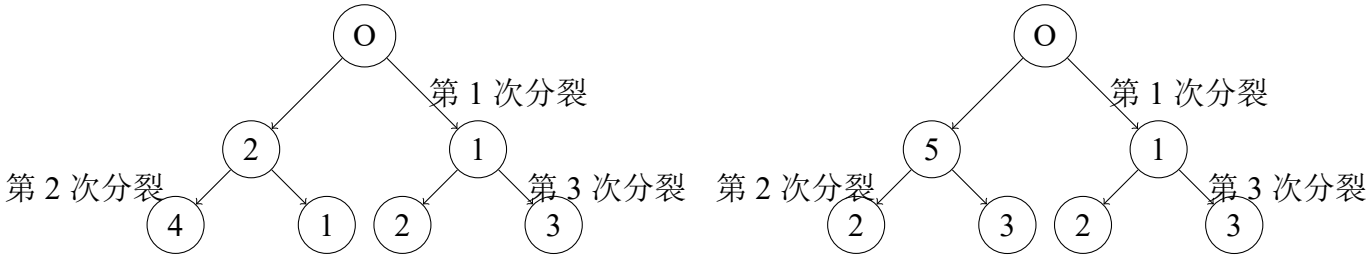


图 4　计算特征重要度

5.3.2 lightGBM

GBDT (Gradient Boosting Decision Tree) [4] 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不

易过拟合等优点。GBDT 在工业界应用广泛，通常被用于点击率预测，搜索排序等任务。

LightGBM（Light Gradient Boosting Machine）是微软开源的一个实现 GBDT 算法的框架，支持高效率的并行训练。LightGBM 的提出让 GBDT 在海量数据的情况下有了更高效的实现，让 GBDT 在生产环境中有更好的实践效果，其具有以下优点：

- 训练速度提升
- 内存占用降低
- 准确率提升
- 并行化学习的支持
- 海量数据处理支持

相较于梯度提升决策树模型 XGBoost[5]，LightGBM 拥有近 10 倍的速度以及 1/6 的内存占用，甚至准确率也更高。

5.3.3 方差选择法

方差选择法作为特征筛选的一个指标，首先计算各个特征的方差，之后根据阈值，选择方差大于阈值的特征。因为方差较小的特征往往变化很小，对最终的结果影响相对较小，可以次要考虑。利用 skit-learn 机器学习库，很容易实现删除较小方差特征，即所有样本中具有大部分相同值的特征。

5.3.4 基于树的特征选择

基于树的估计器 [6]，可用于计算特征重要性，也可用于丢弃不想管的特征。

extra Trees 是 Extremely Randomized Trees 的缩写，这是一种组合方法，类似随机森林。对于普通决策树，每个特征都是根据某个标准（信息增益等）去进行划分。而对于 extra trees 中的决策树，划分点的选择更为随机，随机选择一个划分点，然后再按照评判标准选择一个特征。可以保证选出的特征较为重要。

5.4 第一轮特征筛选

首先，对最终要建模的目标变量 CP_S、RON_{loss} 进行相关性分析，结果如表3。

表 3　CP_S 和 RON_{loss} 相关性分析结果

	RON_{loss}	CP_S
RON_{loss}	1	-0.24
CP_S	-0.24	1

可得，两变量相关性较低，于是我们对 CP_S、RON_{loss} 分别分析。

为了对 CP_S 以及 RON_{loss} 的整体分布有一个更清晰的概念，我们将数据集中 RON_{loss}，CP_S 的数值分布频率进行统计，并绘制了数值分布-频数统计图，如图5。

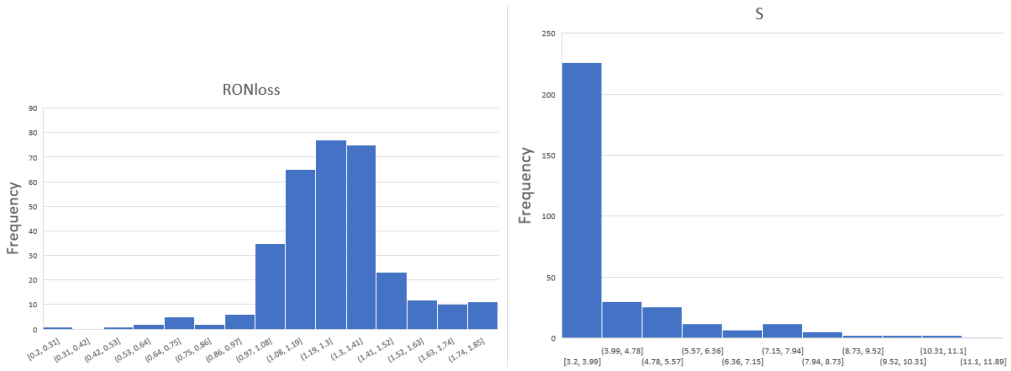


图 5　RON_{loss}、CP_S 数值分布-频数统计图

可以观察到，RON_{loss} 的值分布较为均匀，在 $[0.97, 1.85]$ 区间上都有分布。但是 CP_S 的分布绝大多数集中在 $[3.2, 2.99]$ 区间。

由于 CP_S 数值分布的不均匀性，可能会对模型学习结果的准确性产生一定的影响。

对已经进行最大最小限幅以及 3σ 准则筛选过后的数据，针对 CP_S、RON_{loss} 利用 lightGBM 做特征权重打分。各筛选出排名前 40 的较重要变量，如图6、7所示。

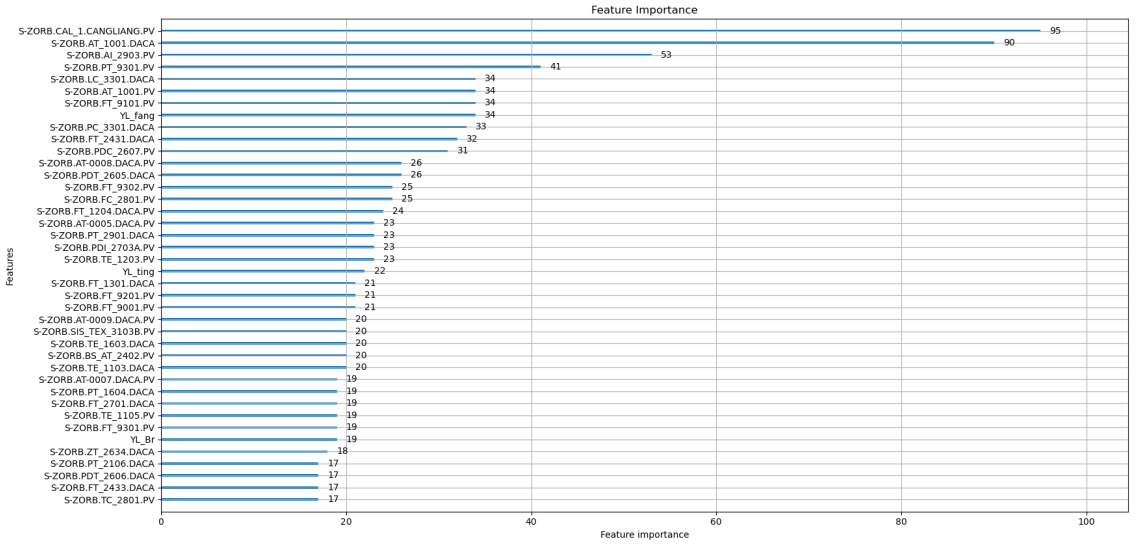


图 6　基于 CP_S 特征权重排名前 40 的操作变量

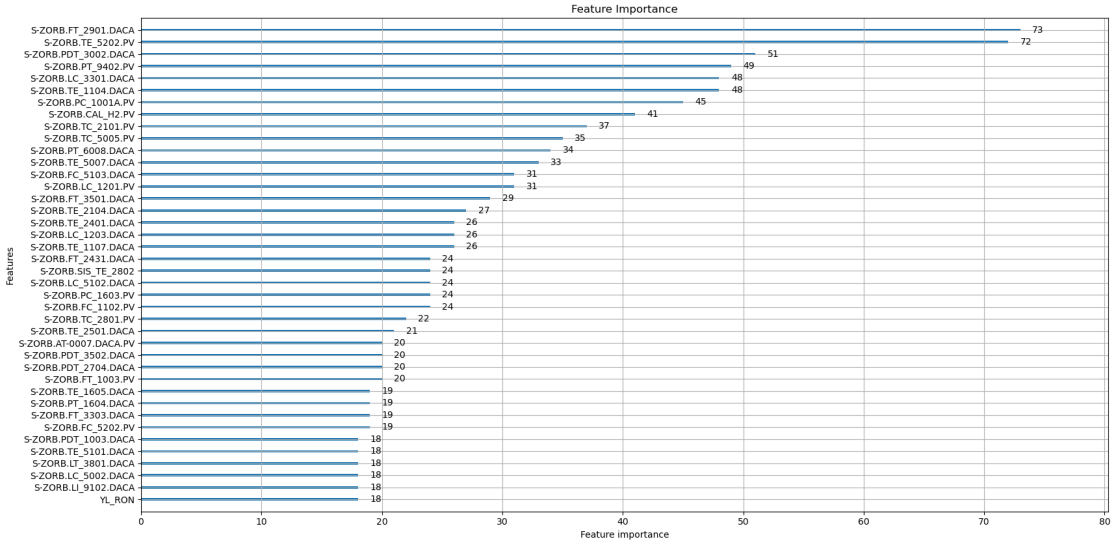


图 7　基于 RON_{loss} 特征权重排名前 40 的操作变量

其中都出现的变量以及基于 CP_S、RON_{loss} 得到的特征权重分别排名前五的变量如表4。

表 4　特征权重筛选出的操作变量

按 CP_S 特征权重前 5		按 RON_{loss} 特征权重前 5	
列号	操作变量名	列号	操作变量名
23	$S-ZORB.PT\_9301.PV$	20	$S-ZORB.TE\_5202.PV$
186	$S-ZORB.AT\_1001.DACA$	40	$S-ZORB.PT\_9402.PV$
40	$S-ZORB.PT\_9402.PV$	136	$S-ZORB.LC\_3301.DACA$
341	$S-ZORB.CAL.CANGLIANG.PV$	151	$S-ZORB.FT\_2901.DACA$
136	$S-ZORB.LC\_3301.DACA$	252	$S-ZORB.PDT\_3002.DACA$

CP_S，RON_{loss} 前 40 特征权重共同出现的操作变量	
列号	操作变量名
76	$S-ZORB.TC\_2801.PV$
136	$S-ZORB.LC\_3301.DACA$
175	$S-ZORB.FT\_2431.DACA$
285	$S-ZORB.PT\_1604.DACA$

基于此题背景，原料性质 YL_{fang}、YL_{ting}、YL_{Br}、YL_{RON} 均保留。由于原料性质的重要程度极高，待生吸附剂性质、再生吸附剂性质中 DS_S、DS_C、ZS_S、ZS_C 也纳入主要变量。

如此，挑选出共计 19 个变量。

针对 19 个变量做相关性分析，绘制热图，发现吸附剂性质中 DS_S 和 DS_C、ZS_S 和 ZS_C 相关性较高，如图8所示，因此再生吸附剂性质与待生吸附剂性质中各剔除一个。

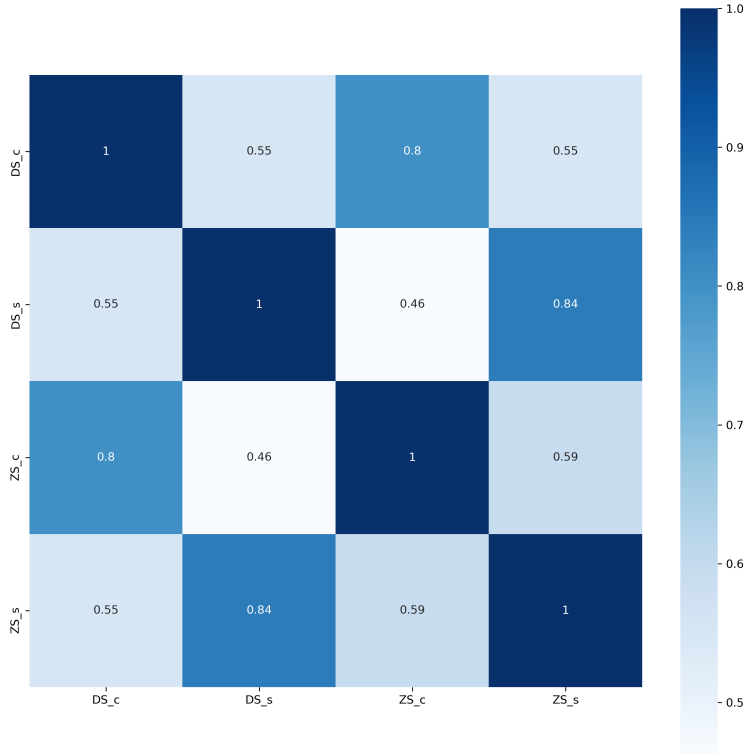


图 8　19 个变量的相关性热图

经过第一轮筛选，共保留 17 个变量。

5.5 第二轮特征筛选

针对排名较低且各自独立出现的变量，共计 57 个，进一步利用相关性分析方法针对变量之间相关性 >0.8 的两两变量，判定为较强相关性，避免同时出现在选择的主要变量中。

同时，利用基于决策树的模型再做特征选择，优中选优，得到如表5所示 5 个操作变量。

表 5　决策树模型再次选择后的五个操作变量

列号	操作变量名
55	$S-ZORB.TE\_1105.PV$
35	$S-ZORB.FT\_9201.PV$
246	$S-ZORB.PDT\_1003.DACA$
229	$S-ZORB.ZT\_2634.DACA$
195	$S-ZORB.FC\_5103.DACA$

5.6 特征筛选结果

经过筛选流程，我们获得了包括原料性质和吸附剂性质在内的 22 个主要变量，如表6所示。

表 6　筛选出来的特征

变量性质	变量名称
吸附剂性质	DS_C
吸附剂性质	ZS_S
原料性质	YL_{fang}
原料性质	YL_{Br}
原料性质	YL_{ting}
原料性质	YL_{RON}

列号	变量名称
20	$S-ZORB.TE\_5202.PV$
23	$S-ZORB.PT\_9301.PV$
35	$S-ZORB.FT\_9201.PV$
40	$S-ZORB.PT\_9402.PV$
55	$S-ZORB.TE\_1105.PV$
76	$S-ZORB.TC\_2801.PV$
136	$S-ZORB.LC\_3301.DACA$
151	$S-ZORB.FT\_2901.DACA$
175	$S-ZORB.FT\_2431.DACA$
186	$S-ZORB.AT\_1001.DACA$
195	$S-ZORB.FC\_5103.DACA$
229	$S-ZORB.ZT\_2634.DACA$
246	$S-ZORB.PDT\_1003.DACA$
252	$S-ZORB.PDT\_3002.DACA$
285	$S-ZORB.PT\_1604.DACA$
341	$S-ZORB.CAL.CANGLIANG.PV$

6. 问题三的分析与求解

6.1 集成学习

集成学习（Ensemble Learning）是一种能在各种的机器学习任务上提高准确率的强有力技术，其通过组合多个基学习器来完成学习任务。基学习器一般采用的是弱可学习器，通过集成学习，组合成一个强可学习器，如图9。

弱可学习，是指学习的正确率仅略优于随机猜测的多项式学习算法；强可学习指正确率较高的多项式学习算法。

集成学习的泛化能力一般比单一的基学习器要好。

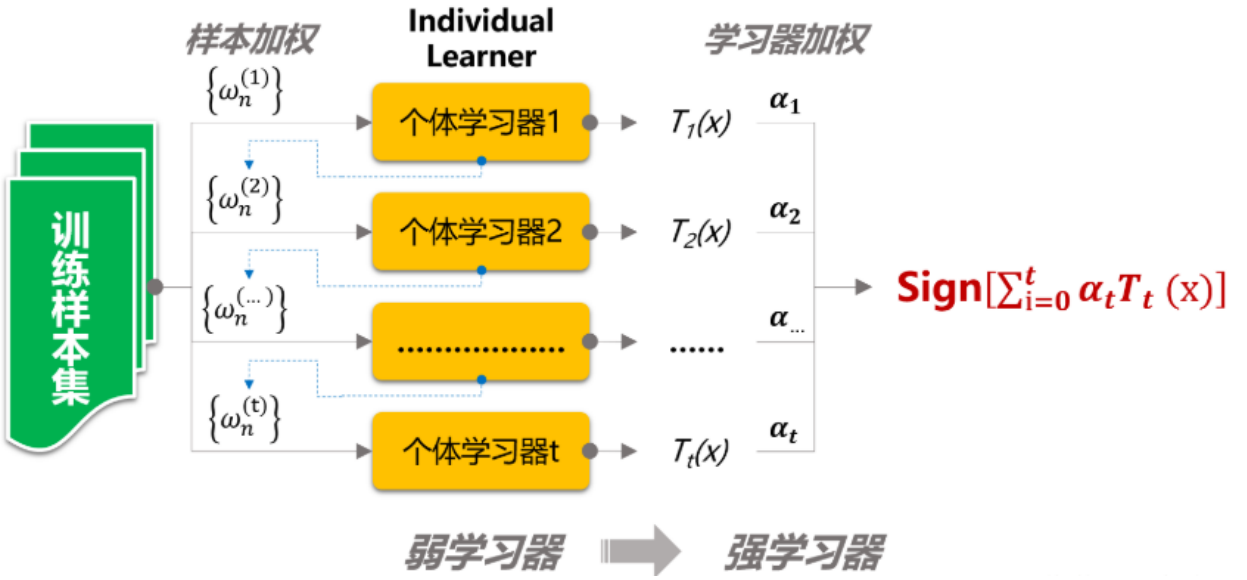


图 9　集成学习示意图

6.2 评价指标

无论利用机器学习算法进行回归、分类或者聚类时，评价指标，即检验机器学习模型效果的定量指标，都是一个不可避免且十分重要的问题。

对于回归问题，本文采用表7所示的两种评价指标。

表 7　回归问题评价指标

全称	简写	含义
mean squared error	MSE	均方差
mear absolute error	MAE	平均绝对误差

均方误差（MSE）是指参数估计值与参数真实值之差平方的期望值，反应自变量与因变量之间的相关程度。MSE 可以评价数据的变化程度，MSE 的值越小，说明预测模型描述实验数据具有更好的精确度。

平均绝对误差（MAE）是绝对误差的平均值，能更好地反映预测值误差的实际情况，同时对异常点的检测非常有效。

6.3 基学习器

根据问题二筛选出的建模主要变量，利用一系列基学习器对 RON_{loss}，CP_S 进行分别建模预测。

其中基学习器采用如下几种方案：

决策树（dTree）是一个预测模型，代表的是对象属性与对象值之间的一种映射关系。

随机森林（Random Forest）是利用多棵决策树对样本进行训练并预测的模型。

梯度提升决策树（lightGBM）是在决策树做运算时，对其梯度进行提升。

多层感知机（MLP）是一种前馈人工神经网络模型，其将输入的多个数据集映射到单一的输出的数据集上。

对上述 4 个基学习器单独训练之后，做各基学习器相关性热图10所示，除个别基学习器相关性较强以外，大部分具有较弱相关性，利于模型融合与集成。

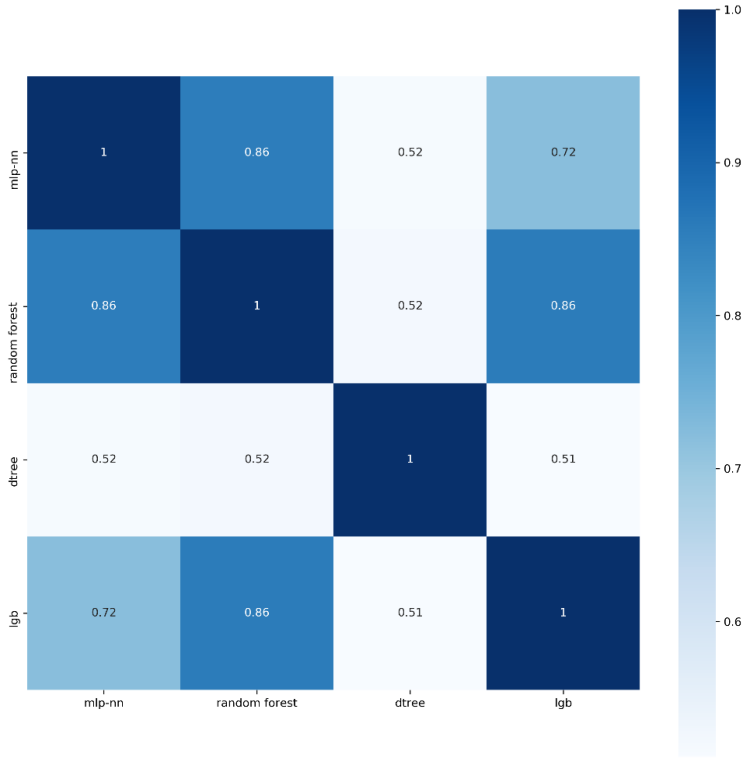


图 10　各基学习器相关性热图

6.4 元学习器

根据上文训练好的基学习器，将预测偏差再次传递给元学习器：梯度下降树（GBDT），做模型融合的集成学习。

GBDT[7] 是一种 boosting 算法，该算法由多棵 CART 回归树组成，所有树的结论累加起来做最终答案。简言之是在拟合上一个模型产生的残差，最终将多棵树的预测结果相加。

6.5 预测结果分析

对于目标其一：RON_{loss}，将集成学习的结果与基学习器结果对比，如图11所示，点越靠近虚线表示预测值越接近真实值，集成学习效果明显，预测较为精准。除个别偏差外，预测值和真实值较为接近。

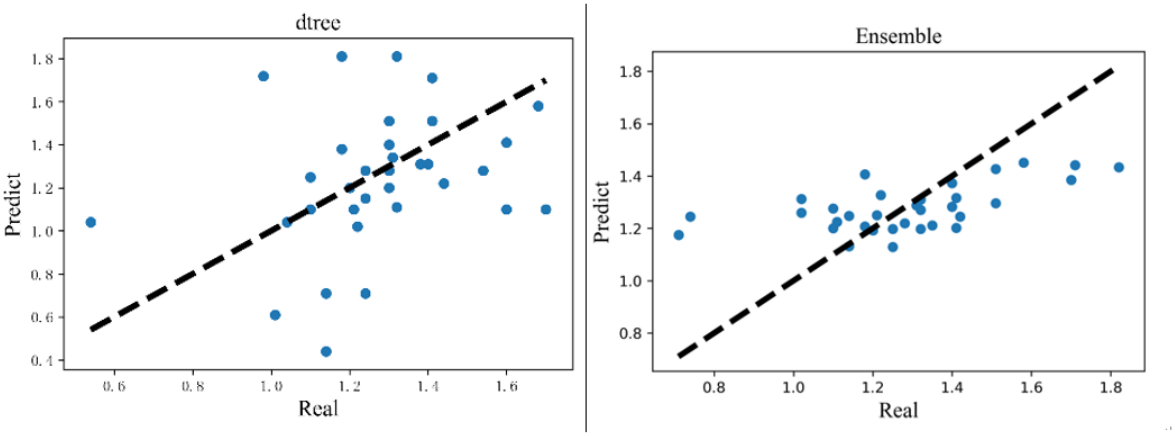


图 11　RON_{loss} 集成学习与基学习器对比

且根据评价指标表8可以看出模型融合后的结果相较于各个基学习器的误差都较小，预测效果明显

表 8　RON_{loss} 评价指标 MAE、MSE 对比表

评价指标 S	MSE	MAE
$mlpnn$	0.021	0.107
$randomforeast$	0.02	0.112
XGB	0.018	0.099
$dtree$	0.042	0.155
lgb	0.018	0.102
$ensnmble$	0.01	0.092

对于目标其一：CP_S，将集成学习的结果与基学习器结果对比，如图12，点越靠近虚线表示预测值越接近真实值，集成学习效果一般，但除个别离散点有较大偏差外，预测值和真实值较为接近。

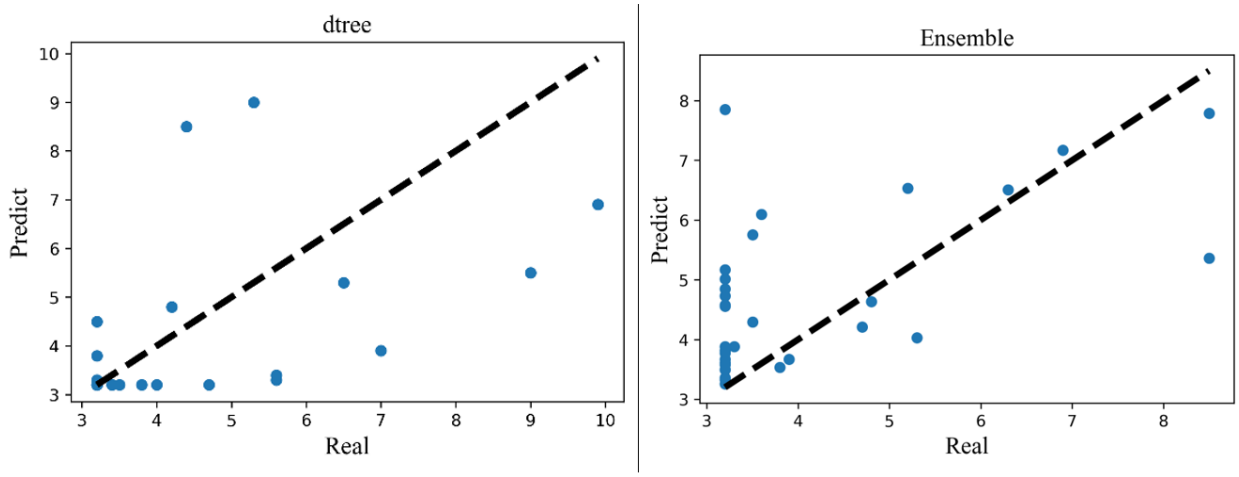


图 12　CP_S 集成学习与基学习器对比

根据评价指标表9可以看出模型融合后的结果相较于各个基学习器的相对误差都较小，集成效果较好。

表 9　CP_S 评价指标 MAE、MSE 对比表

评价指标 S	MSE	MAE
$mlpnn$	0.025	0.115
$randomforeast$	0.031	0.128
XGB	0.027	0.116
$dtree$	0.051	0.15
lgb	0.028	0.11
$ensnmble$	0.022	0.095

综合分析，集成学习对辛烷的损失值 RON_{loss}、产品的含硫量 CP_S 预测结果较好。

7. 问题四的分析与求解

7.1 建模分析

根据题意，需要在保持原料、待生吸附剂、再生吸附剂的性质不变的情况下，优化操作变量，使得产品的硫（CP_S）含量不大于 5$\mu g/g$，并且产品的辛烷值损失降幅大于 30%。

在第三问中，根据相关性分析，已经建立了基于同样的特征（操作变量），分开预测产品的硫含量（CP_S）与辛烷值的损失值（RON_{loss}）的集成学习模型，并且经过误差的校验，是一个非常正确的预测模型。由此获得了两个机器学习模型 $Model_{RON_{loss}}$，$Model_{CP_S}$，如图13所示。

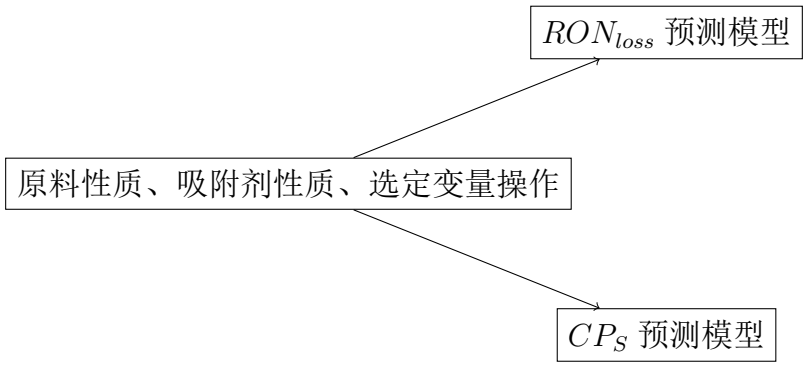


图 13　通过主要变量求得两个目标的预测模型

问题转化为在规定范围内找到合理的操作变量取值,同时使得由 $Model_{RON_{loss}}, Model_{CP_S}$ 预测的 RON_{loss} 和 CP_S 满足题目要求的条件。

由于不一定 325 个数据样本中每个样本都能在规定范围内找到适合的操作变量取值，故首先计算满足产品硫含量与辛烷值要求范围的操作变量最优取值，再从 325 个样本中筛选出真正满足规定范围的样本，以及其对应操作变量。

7.2 优化方程

经过分析，本文转化为为非传统多目标优化问题。

对每一个样本求解如下优化方程：

$$min\{Z_1 = F_1(x_1), Z_2 = F_2(x_2)\}, St.low \leq x_i \leq high$$

其中，x_i 代表原料性质、吸附剂性质以及选定操作变量取值。Z_1 为 RON_{loss}、Z_2 为 CP_S。

7.3 遗传算法

与传统的多目标优化问题不同，无法得知目标函数的具体表达，目标函数只是由机器学习建立的一个黑盒映射，无法利用梯度下降等需要知道函数表达的优化方式。

遗传算法是一种智能算法，用于多目标优化时，只需计算目标函数的值，对优化问题本身的性质要求非常低，尤其区别于数学优化算法（例如梯度下降法）往往依赖于一大堆的条件，如函数是否为凸优化，目标函数是否可微等等。

将遗传算法作为我们的优化算法，将机器学习模型作为我们的目标函数，将得到令人满意的结果。

常见的遗传算法流程如下：

（1）初始化规模为 N 的种群，其中染色体每个基因的值采用随机数产生器生成并满足问题定义的范围。当前进化代数 $Generation = 0$。

（2）用评估函数对种群中所有染色体进行评价，分别计算每个染色体的适应值, 保存适应值最大的染色体 $Best$。

（3）采用轮盘赌选择运算对种群的染色体进行选择操作，产生规模同样为 N 的种群。

（4）按照概率 P_c 从种群中选择染色体进行交叉运算。两两父代染色体交换部分基因，产生两个新的子代染色体，子代染色体取代父代染色体进入新种群。没有进行交叉的染色体直接复制进入新种群。

（5）按照概率 P_m 对新种群中染色体的基因进行变异操作。发生变异的基因数值发生改变。变异后的染色体取代原有染色体进入新种群，未发生变异的染色体直接进入新群体。

（6）变异后的新种群取代原有种群，重新计算种群中各个染色体的适应值。倘若种群的最大适应值大于 $Best$ 的适应值，则以该最大适应值对应的染色体替代 $Best$ ，即更新最大适应值大于 $Best$ 。

（7）当前进化代数 $Generation$ 加1。如果 $Generation$ 超过规定的最大进化代数或 $Best$ 达到规定的误差要求，算法结束，$Best$ 可表示问题的一个解; 否则返回（3）。

遗传算法流程如图14。

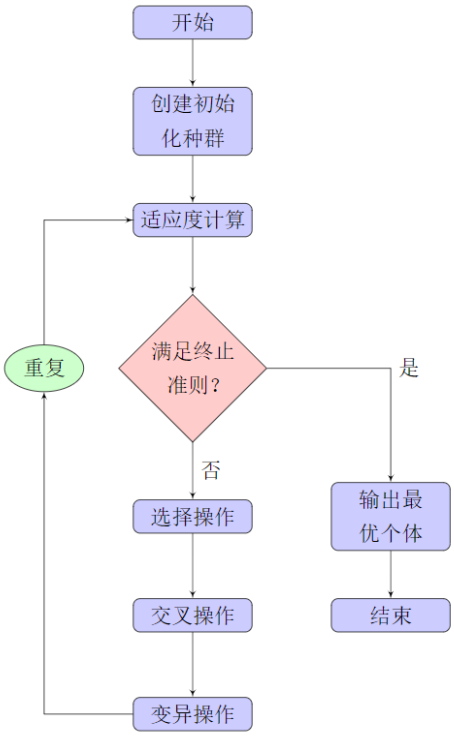


图 14　遗传算法流程图

由于建模问题为多目标优化，本文采取的遗传算法主要基于 NSGA-II[8]，通过拥挤度来度量系统元素分布的情况，以此来判断解的适应度，从而选择出那些分布均匀，获得信息最多的基因。从后面的例子中可以看出，我们的程序可以在进化代数不多的情况下，很快地求得我们所需的帕累托最优解。

7.4 帕累托最优解

多目标优化未必存在一个极值点，同时使两个目标函数达到最小值，于是将极值点转化为帕累托最优解。

帕累托最优（Pareto Optimality）[9]，是指资源分配的一种理想状态。假定固有的一群人和可分配的资源，从一种分配状态到另一种状态的变化中，在没有使任何人境况变坏的前提下，使得至少一个人变得更好。帕累托最优状态就是不可能再有更多的帕累托改进的余地；换句话说，帕累托改进是达到帕累托最优的路径和方法。

就多目标优化问题而言，对本题两个目标函数，如果存在解 A，在变量空间中找不到其他的解能够优于解 A（注意这里的优于一定要两个目标函数值都优于 A 对应的函数值），那么解 A 就是帕累托最优解。

帕累托最优解只是问题的一个可接受解，一般都存在多个帕累托最优解。对于本题而

言，如果我们能找到一个符合限制条件的支配点，则此点将是使得问题最优化的取值。然而现实情况是，当我们想要最小化产品辛烷值的损失时，产品硫含量可能达不到最小值，甚至可能超出对产品硫含量的限制，当我们想要减少产品硫含量的时候，产品的辛烷值损失可能不能接受。所以我们将依据帕累托最优的准则，挑选出可能符合条件的操作变量优化值（称为帕累托前沿点），再在后续中进一步筛选出需要的解。

7.5 具体优化过程

在优化程序编写的过程中，我们设置了初始种群规模为 60，每个个体具有一条染色体，最大进化代数为 200 代。以样本号为 67 的样本作为例子，如图所示，在进化的第三代中我们可以看到，初始种群差异较大，且基因不够优秀，并没有分布在 F1,F2 的帕累托最优点附近，在进化的第 12 代，可以看到我们的种群已经尝试地分布在了 F1 或者 F2 的较小取值范围处，在第 27 代时，种群中一些劣势的点已经被淘汰了，保存下来的优秀个体已经分布在问题的帕累托最优解附近，第 110 代时，我们已经找到了问题的两个帕累托最优解。

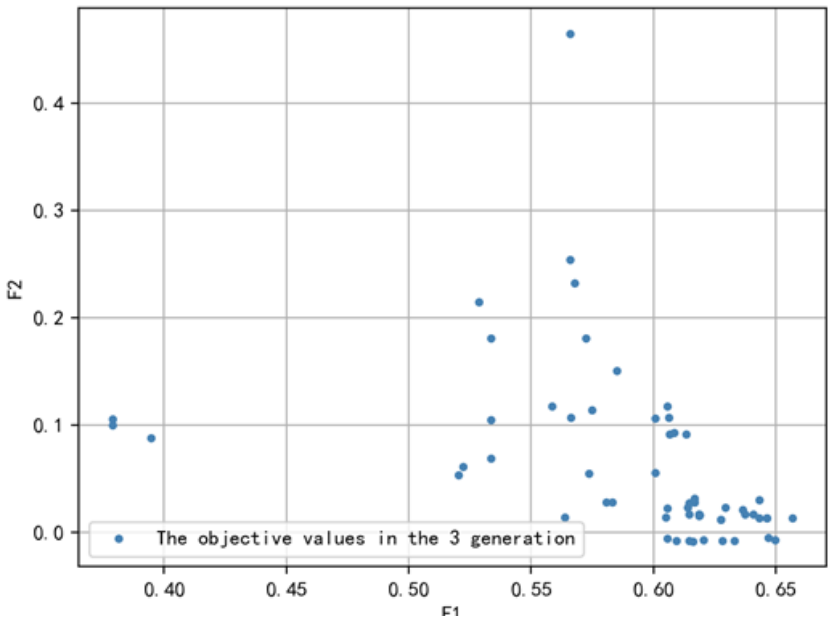


图 15　第 3 代种群变量的目标函数取值

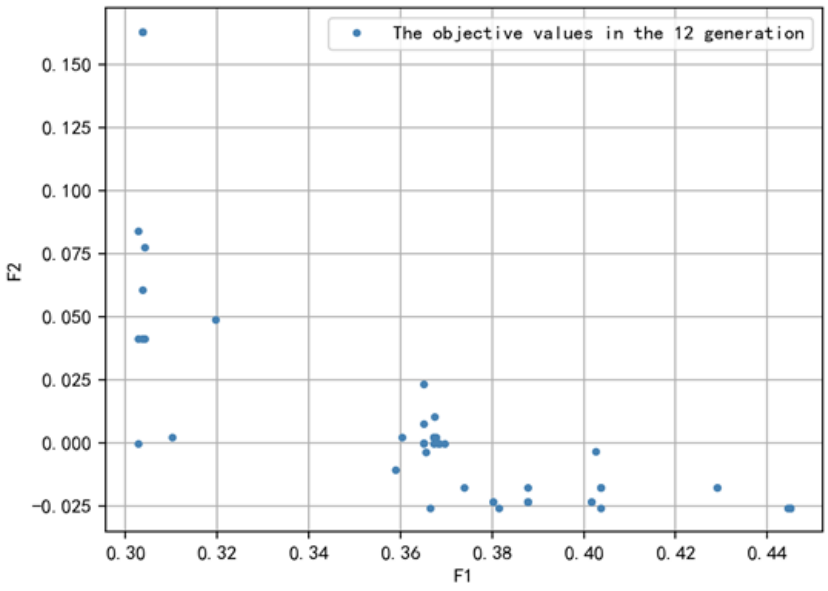


图 16　第 12 代种群变量的目标函数取值

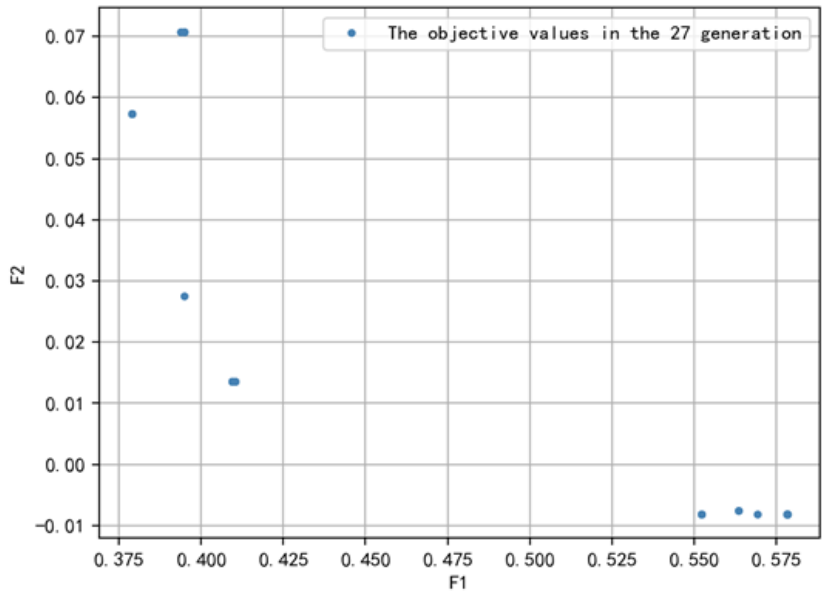


图 17　第 27 代种群变量的目标函数取值

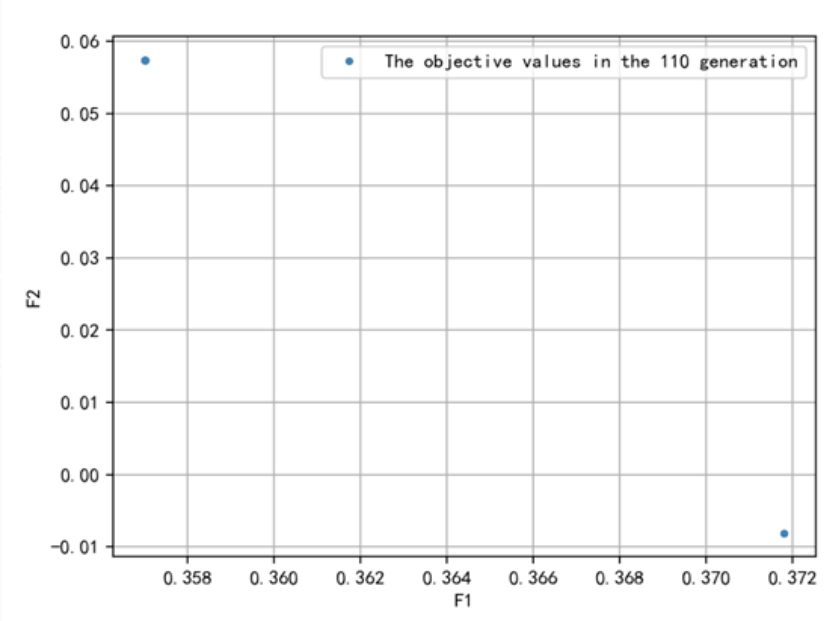


图 18　第 110 代种群变量的目标函数取值

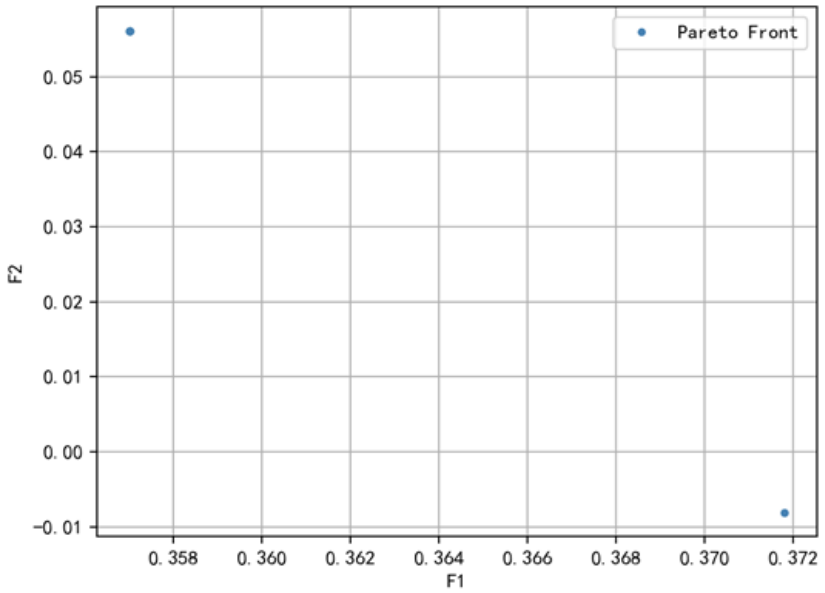


图 19　帕累托最优种群变量目标函数取值

图中 F1 代表产品 RON 损失的取值，F2 代表产品硫含量的取值，这里的取值由于先前经过标准化，与现实取值不同，仅作参考

由于我们只是寻找了 67 号样本的操作变量的帕累托最优，并不保证绝对满足题设要求，再对最优解做筛选，幸运的是，67 号优化所得的帕累托最优解完全满足产品 RON 损

失降低 30%，产品硫含量不大于 $5\mu g/g$ 的要求。

我们找到了两种操作变量的优化可能，具体操作变量取值可见附件 $reslut\_opt.csv$。

- 对于第一种，我们优化的操作变量可以使得 CP_S 取值为 3.01859，RON_{loss} 为 0.71346。
- 对于第二种，我们优化的操作变量可以使得 CP_S 取值为 3.19856，RON_{loss} 取值为 0.42908。

而 67 号样本在没有优化前，CP_S 取值为 3.2，RON_{loss} 取值为 1.2800000000000011。

若采取第一种优化方式，我们可以使得 RON_{loss} 减小 44.26%，但有较小的 CP_S，若采取第二种优化方式，我们可以使得 RON_{loss} 减小 66.48%，但有较大的 CP_S ，由此，我们便可以依据实际生产的要求选择合适的优化方案

7.6 结果总结

最终，我们找到了所有样本的帕累托最优点，并且一些样本有不止一个帕累托最优点。一些样本中甚至还找到了 7 个最优点，说明操作变量的优化还是有很多空间的，样本数与帕累托最优解数量如表10。

表 10　帕累托最优值与样本数对应关系

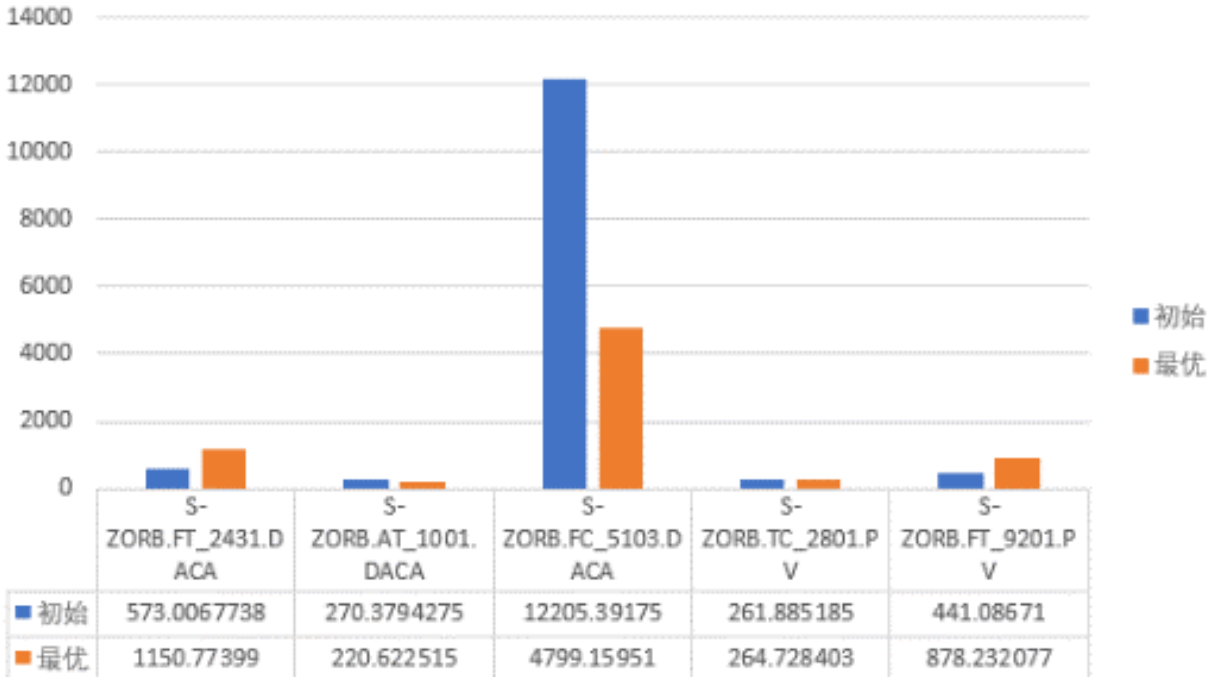
样本数	帕累托最优解数量
77	1
93	2
79	3
47	4
19	5
8	6
2	7

再依据题设条件限制，在 325 个样本中，我们发现了其中有 304 个样本存在可以使得限制条件满足的优化操作变量的方式，其中一些样本同时具有多种优化方式，基于对 RON_{loss} 或者 CP_S 的侧重，我们可以选择不同的优化后的操作变量，具体优化后的操作变量取值以及基于优化后的操作变量可以获得的 RON_{loss} 以及 CP_S，见附件 $result\_opt.csv$，其中对于每一个样本给出了详尽的数据。

8. 问题五的分析与求解

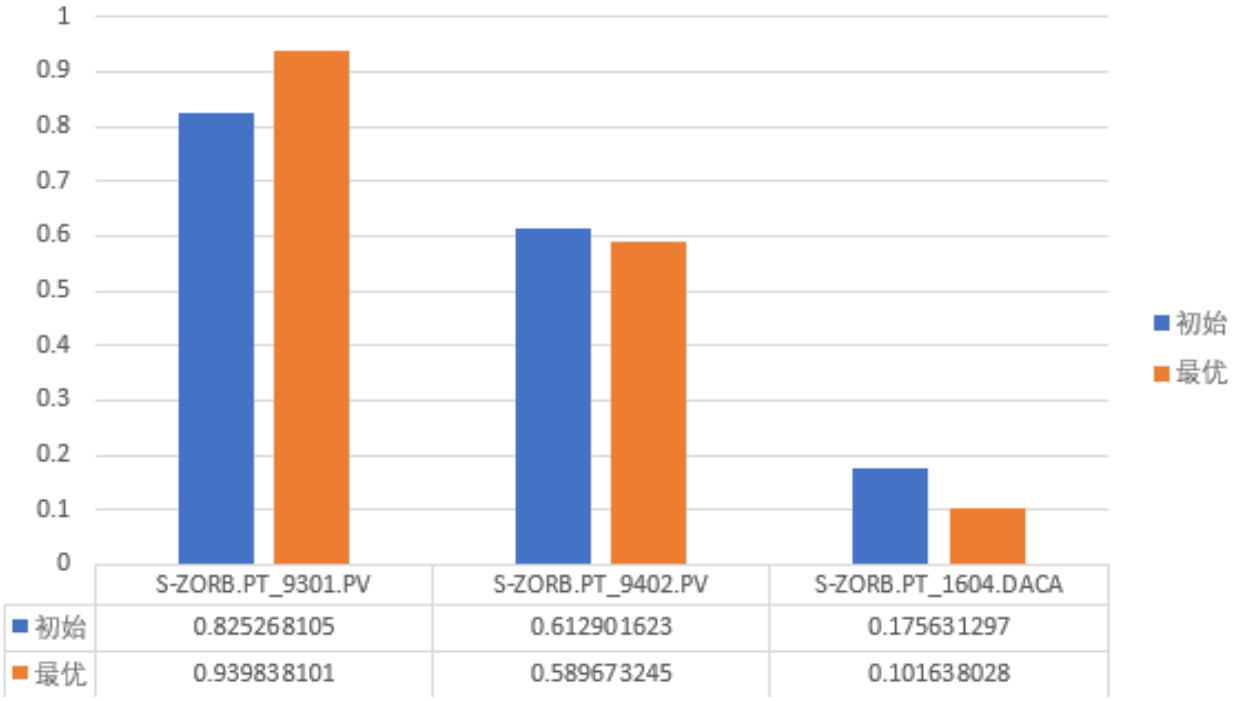
8.1 133 号样本最优解

根据问题四中构建的优化策略，在保证优化目标的前提下，控制 133 号样本点的原料性质、待生吸附剂和再生吸附剂的性质数据保持不变，寻找帕累托最优解对应的操作变量，与初始操作变量对比，结果如图20-22。



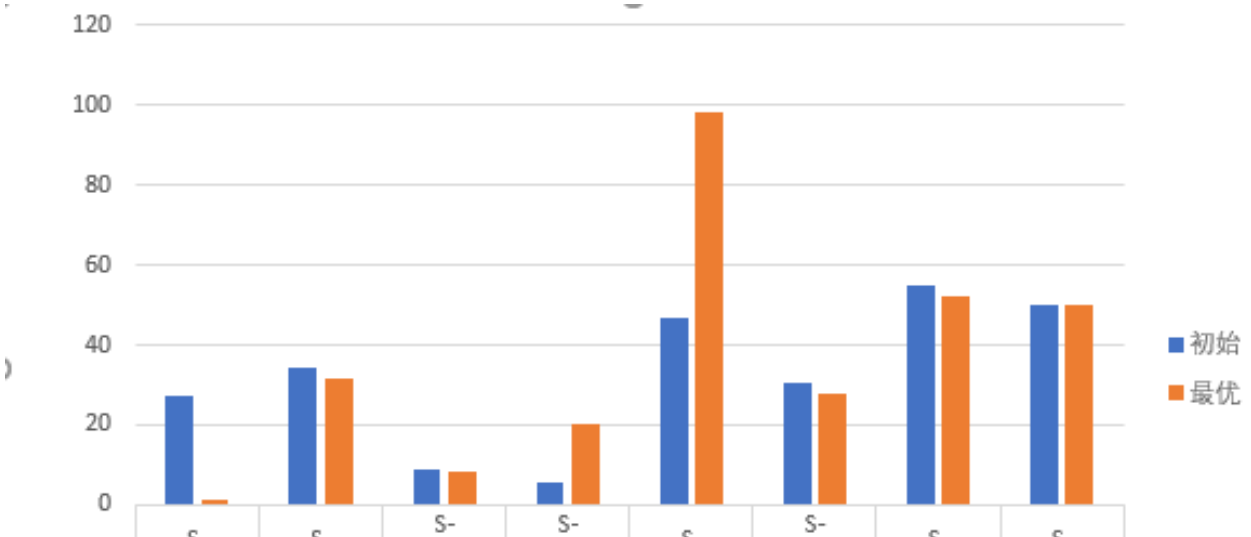
	S-ZORB.FT_2431.DACA	S-ZORB.AT_1001.DACA	S-ZORB.FC_5103.DACA	S-ZORB.TC_2801.PV	S-ZORB.FT_9201.PV
初始	573.0067738	270.3794275	12205.39175	261.885185	441.08671
最优	1150.77399	220.622515	4799.15951	264.728403	878.232077

图 20　最优解对应操作变量与初始值的对比（1）



	S-ZORB.PT_9301.PV	S-ZORB.PT_9402.PV	S-ZORB.PT_1604.DACA
初始	0.825268105	0.612901623	0.175631297
最优	0.939838101	0.589673245	0.101638028

图 21　最优解对应操作变量与初始值的对比（2）



	S-ZORB.FT_2901.DACA	S-ZORB.TE_5202.PV	S-ZORB.PDT_1003.DACA	S-ZORB.PDT_3002.DACA	S-ZORB.ZT_2634.DACA	S-ZORB.CAL.CANGLIANG.PV	S-ZORB.TE_1105.PV	S-ZORB.LC_3301.DACA
初始	27.443595	34.452484	8.84629	5.8075056	46.813061	30.730875	54.788542	50.173972
最优	1.1120726	31.843387	8.4985492	20.018086	98.166755	27.65025	52.304558	50.195204

图 22　最优解对应操作变量与初始值的对比（3）

其中 $S-ZORB.FC\_5103.DACA$、$S-ZORB.ZT\_2634.DACA$ 两个变量初始值与优

化值相差较大，说明其对最终目标影响性较高，后面工程实际应该重点考虑。

图23展示了 CP_S、RON_{loss} 与初始数据样本中的对比。

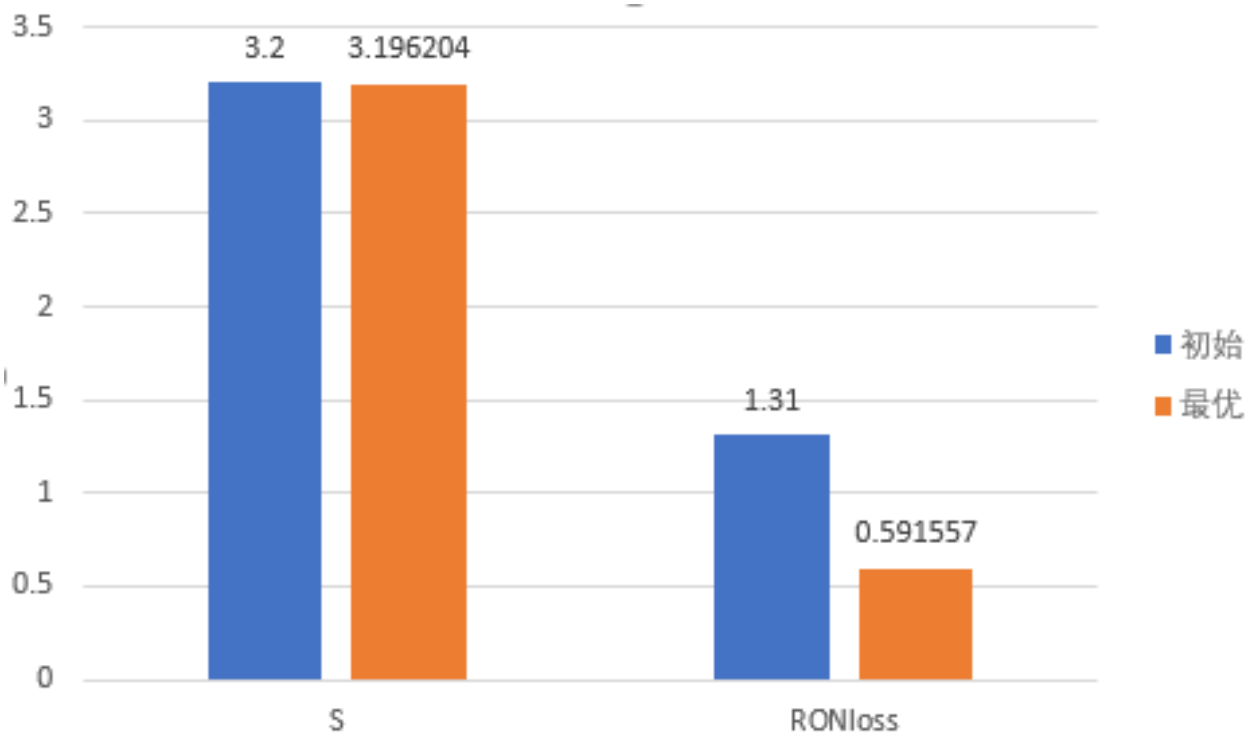


图 23　优化后的 CP_S、RON_{loss} 与初始的 CP_S、RON_{loss} 对比

可以看出 CP_S 保持稳定，含量相对较小，而 RON_{loss} 降幅达到 54.84%，优化效果显著。

8.2 调整策略

根据优化结果，结合各主要操作变量每次允许的调整幅度，构建如下操作变量调整策略：

- 从初始值开始，使各主要操作变量每次都向理想值 +/-最大调整幅度。
- 若某操作变量提前达到最优数值，则后续调整步保持最优数值，其余操作变量继续调整，直到每个操作变量实际值等于最优操作变量数值

通过计算，绘制出各操作变量的调整次数分布图，如图24所示，操作变量中的 $S-ZORB.PDT\_3002.DACA$、$S-ZORB.ZT\_2634.DACA$ 分别需要调整 15、11 次才能达到最优变量值，说明工程实际中仍具有较大改进范围。

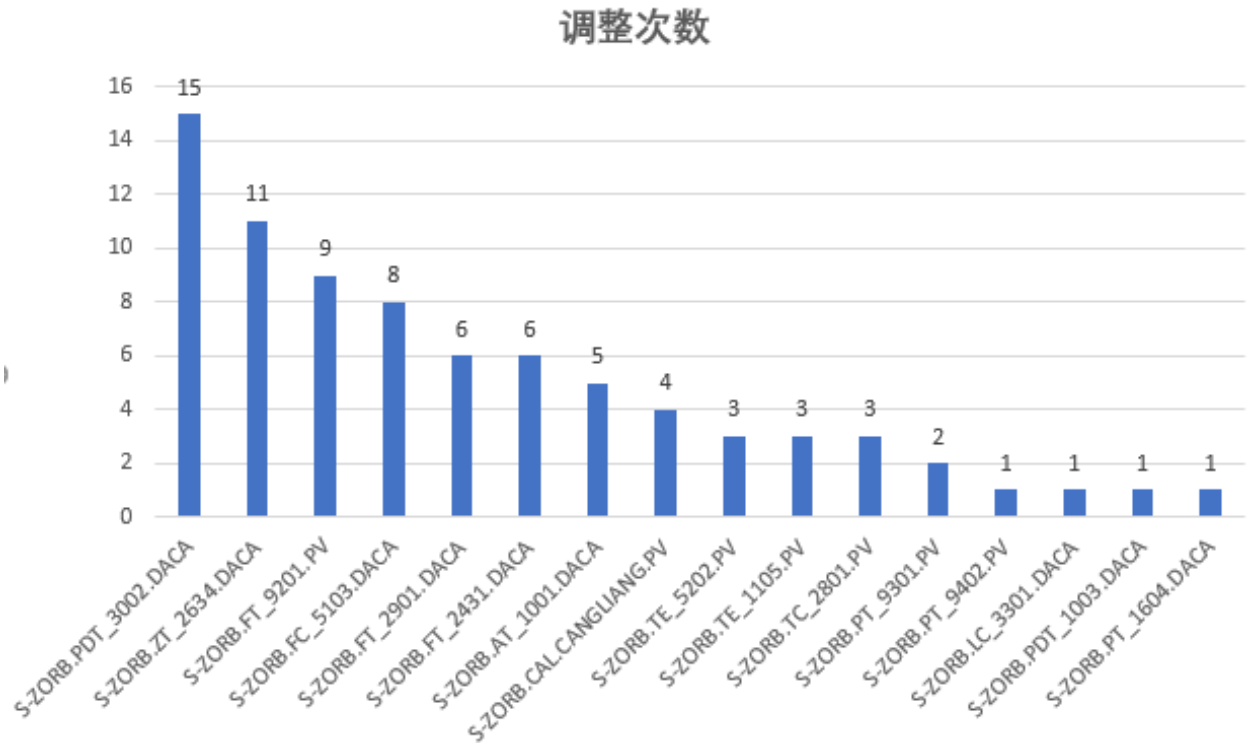


图 24　各操作变量的调整次数分布

8.3 优化预测

根据调整策略，每个调整步形成一个新的样本，共计 15 个样本，利用问题三中构建的集成学习模型预测优化过程中对应的 CP_S、RON_{loss}。将其变化过程分别绘制，如图25、26。

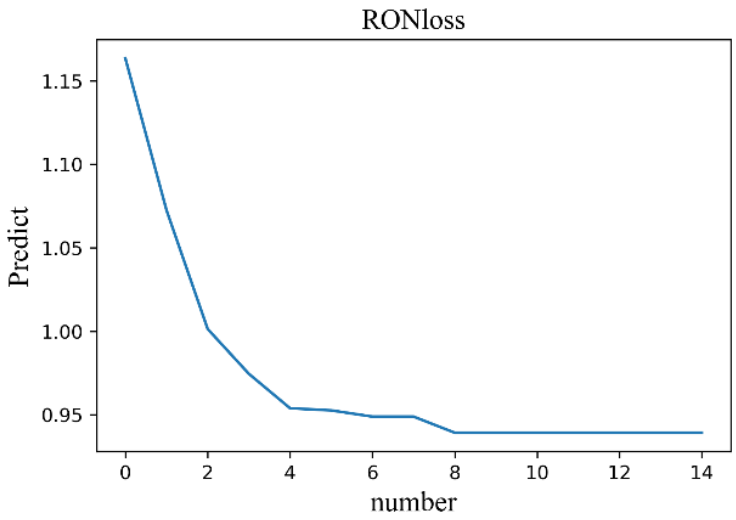


图 25　RON_{loss} 预测优化过程中的变化

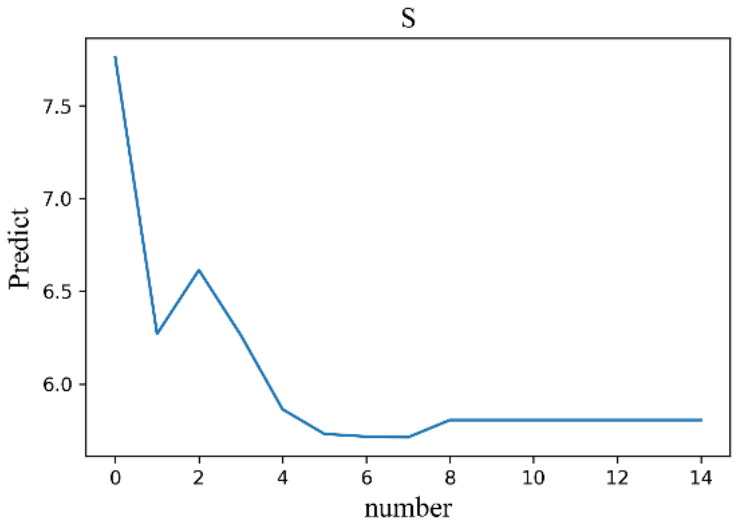


图 26　CP_S 预测优化过程中的变化

随着操作变量逐渐趋向最优操作变量，RON_{loss} 逐渐降低，且在优化初期由于多个变量同时更改，RON_{loss} 的下降非常明显，证明优化策略有显著效果。

操作变量优化的同时，CP_S 也出现了降低，但由于 CP_S 该变量的分布，导致 CP_S 的预测模型误差较大，具体数值不具有较大参考意义，但仍能很明显看到优化策略可以一定程度降低 CP_S。

8.4 操作变量与目标变化趋势

将上文中分析具有重要地位的三个操作变量：$S-ZORB.PDT\_3002.DACA$、$S-ZORB.ZT\_2634.DACA$、$S-ZORB.FC\_5103.DACA$ 及其调整过程中对应的 RON_{loss} 和 CP_S 的变化轨迹绘制如图27。

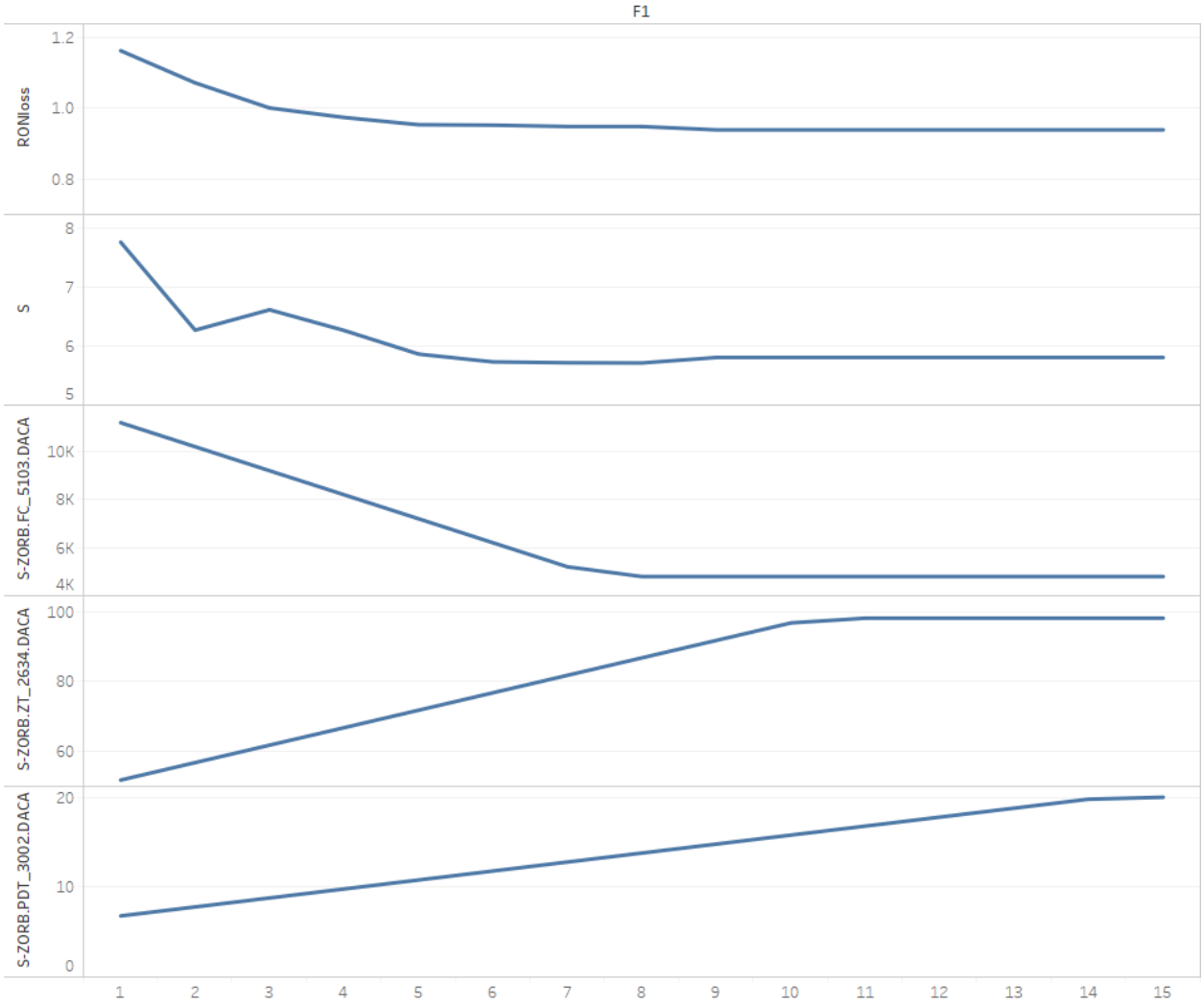


图 27　三个重要操作变量及其调整过程中对应的 RON_{loss} 和 CP_S 的变化

可以发现随着 $S-ZORB.FC\_5103.DACA$ 降低、$S-ZORB.PDT\_3002.DACA$ 和 $S-ZORB.ZT\_2634.DACA$ 不断增加，RON_{loss} 和 CP_S 逐渐降低。

9. 总结

9.1 模型评价

本文结合数据规则，对数据样本进行清洗，进而利用机器学习算法筛选特征，依据数据样本及其操作变量，构建集成学习来建立 RON_{loss} 的预测模型，模型预测效果较好，误差评价标准较低。为满足目标条件, 进一步利用遗传算法优化预测模型，搜寻帕累托最优解，成功给出每个样本的优化操作条件。针对题目指定的 133 号样本，深入研究了影响其 RON_{loss}、CP_S 的特征，发现具有指导意义的 3 个重要特征。

模型的巨大优势在于，对于一个复杂的现实工业问题，可以筛选出影响目标的重要操作变量，在满足约束条件和优化目标的前提下，进一步优化操作变量，给提高生产力提供了精确的指导意见。

具体而言，利用遗传算法进行优化，可以处理任意复杂的函数，且无需函数的具体表达式，随着进化代数的递增，算法总能找到约束条件下的最优值。但遗传算法是一种带有随机性的优化算法，并非每次的最优点就是全局最优解，其最终结果受到初始种群分布的影响。

9.2 模型改进的思路

1、特征筛选时可以结合生产实际，使模型更具有现实性。

2、预测模型中可以增加逻辑回归等传统算法作为基学习器。

参考文献

[1] 马孝腊, 烟叶分级中若干特征筛选方法的研究.

[2] 张丽新, 高维数据的特征选择及基于特征选择的集成学习研究.

[3] 如何进行特征选择实践, https://blog.csdn.net/u010899985/article/details/81699091, 2018-08-15.

[4] [机器学习] 树模型特征重要性原理总结, https://blog.csdn.net/zwqjoy/article/details/97259891, 2019-07-25.

[5] 陆健健, 江开忠, 基于 XGBoost 算法模型的金融客户信用评估研究, 软件导刊, 18(04): 133-136, 2019.

[6] 李国正, 李丹, 集成学习中特征选择技术, 上海大学学报: 自然科学版(05):598-604, 2007.

[7] Sakhnovich A, On the GBDT Version of the Bäcklund-Darboux Transformation and its Applications to Linear and Nonlinear Equations and Weyl Theory, Mathematical Modelling of Natural Phenomena, 5(4):340-389, 2012.

[8] K Deb S A T M, A Pratap, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, 2002.

[9] Ping W, Xianci X, Lemin L, 等, 计及输电阻塞的帕累托最优多目标电网规划, 电子与信息学报, 1995.